在「阻止殺手機器人運動」中，各行各業的專業人士與人道主義者紛紛呼籲禁止自主性武器。圖為2019年3月間在德國柏林舉行的相關運動。(Reuters/達志)

● 作者/Jai Galliott and Jason Scholz ● 譯者/黃國賢 ● 審者/馬浩翔

人工智慧武器化的道德議題

Artificial Intelligence in Weapons: The Moral Imperative for Minimally-Just Autonomy

取材/2018年美國印太事務季刊(*Journal of Indo-Pacific Affairs*, Winter/2018)

人工智慧武器面臨道德與正義的考驗。未來，人工智慧武器是否就此取代戰場上的決策程序，甚或取代指揮官的決策權限？本文從科技、道德及國際法面向檢視可預見的未來。

為建置合法且符合道德正義的軍力，未來的自主性人工智慧系統必不能犯下人道主義錯誤或做出傷及鄰國行為，而為達此一目標，一種符合保護符號、位置及信號受到攻擊的「最低限度正義自主使用人工智慧」(minimally-just autonomy using artificial intelligence, MinAI)預防性舉措就有其發展的必要。「最低限度正義自主使用人工智慧」優於迄今所提出的任一其他符合最大道德正義的形式。本文將探討對人工智慧預期恐懼的心態，如何讓今日武器無法更恪遵國際人道法，其中吾等會特別著重探討1949年8月12日通過的《日內瓦公約》附加協議第36條。[1] 針對論述，批判者可能主張機器學習有可能遭到愚弄，參戰者可做出背信忘義的行為以自保等類似論點。本文會直接切入此議題，包括最近針對人工智慧進行的顛覆性研究，並且結論出讓武器中的「最低限度正義自主使用人工智慧」具備道德意識依然相當重要。

引言

在「阻止殺手機器人運動」(Campaign to Stop Killer Robots)中，一些知名演員、商業領袖、傑出科學家、律師和人道主義者紛紛呼籲禁止自主性武器。[2] 2017年11月2日，該運動團體致信給澳大利亞總理滕博爾(Malcolm Turnbull)，信中提及「澳大利亞人工智慧研究圈正要求總理閣下及政府，讓澳國成為世界第二十個採取全球堅定反對人工智慧武器化立場的國家。」這些支持者因擔憂政府無所作為，遂指出開發自主性武器系統將帶來悲慘的後果：「致命的後果就是機器──而非人類──將決定人的生死。」[3] 他們似乎主張全面禁止人工智慧武器──而這個主張也呼應著渠等對未來世界充斥著微型殺手機器人的想像。[4]

禁止人工智慧武器可能阻絕大家提出現今人道主義危機的種種解決方案。大家每天都能在國際新聞媒體中看到有關傳統武器的報導，試想下列情況：利用從警局偷出的手槍殺害無辜人士，使用來福槍在美國校園犯下大規模槍擊案；在公共場所以車輛來輾殺行人；放置炸彈攻擊宗教場所；以導彈攻擊一列滿載毫無戒心乘客，且正準備通過橋樑的火車；發射飛彈攻擊紅十字會的設施等。隨著人工智慧武器的開發，此類型的攻擊事件將可被避免。這些都是裝載著人工智慧的自主性武器可能介入並拯救性命的真實情況。

對於達到大家期待的非暴力狀態所需之手段感到困惑早已不是什麼新鮮事。在反核行動中，普遍瀰漫著蔑視以簡單科技來達成更為和平解決方法的狀態──貫穿了整個美國與蘇聯對峙的時期，最近又因小型彈頭的發明以及1990年代末期禁止地雷活動而又再度勾起此種蔑視。[5] 而提出以下質疑似乎也合情理：針對為何有武器裝配著先進尋標器，卻無法導入人工智慧來辨識紅十字會標誌，進而中止一道攻擊命令。此外，受保護的重要宗教場所、學校及醫院等地點，位置均可設定在武器中來限制其行動，也能指定未經授權的使用者，無法發射武器攻擊他人。那為何吾等不能開始測試這類創新武器，好讓其得以納入國際武器審查標準？

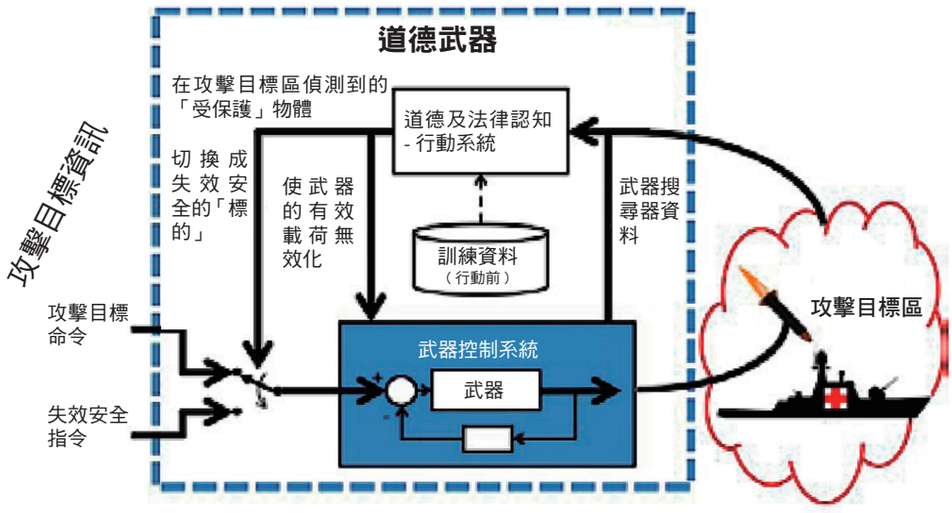
本文斷言，自主化系統在近期也難有能力做出能肩負起道德責任的作為，然而這些系統現在卻

能自主執行那些嵌入其設計與編碼中極富價值的決定，因而得以採取合乎道德與法律標準的行動。[6]

人工智慧的道德範圍

大家必須區分道德能力範圍的兩個末端。透過「最高限度正義自主使用人工智慧」可被接受及無法接受的行動，能夠確保致命行動符合道德義務──即使在當下系統工程師可能無法透過次系統，來辨識出相關致命行動的必要或可行性。然而，符合最大正義的道德機器人需要龐大的道德架構。機器人學家埃金(Ronald Arkin)提出的道德規範，代表符合最高正義系統的原型理論。[7] 道德規範會評估任何提出的致命性行動，使其符合戰爭法及交戰規定。道德規範的關鍵就是限制詮釋，而它的運作說明也顯示了符合最大正義的立場：「限制適用流程應負責導出道德限制，且亦須保證機器人據此產生的行為合乎道德。」[8] 限制系統──以複雜義務及闡述邏輯為基礎──根據同樣複雜的資料結構，來評估系統的戰術推論引擎所產生的行動方案。所謂合乎道德的範圍──正如埃金所述包括比例概念以及交戰規定──往往相當難以界定。

相較之下，雖然「最低限度正義自主使用人工智慧」道德機器人仍是一種僅能限制驅動的系統，但它能在缺乏適當道德規範下，對人類產生的道德行為採取基本抑制，並可相容於多數限制條件中(制式規定而非彈性規定)──意謂所需解讀的系統將會更少。「最低限度正義自主使用人工智慧」處理的是「道德上不被允許」的項目，將限制條件建立在辨識及避開受保護物體和行為等需求上。那些受法律保護的符號及位置、投降信號(包括燈號)，以及無力再戰的地點將被特別避開。大家必須留意到這些人工智慧限制有難易之分，且隨著人工智慧技術的進步而持續改善。「最低限度正義自主使用人工智慧」道德武器的



倘若在有效攻擊目標區內察覺到非預期且受法律道德保護之物體或行為，這類武器能在不遵守攻擊目標命令之情況下，以失效安全的標準執行命令。攻擊目標的資料係由外部取得並回饋武器。

圖1 「最低限度正義自主使用人工智慧」道德武器

谷歌研發的人工智慧系統AlphaGo在擊敗韓國圍棋九段棋手李世乭後，吸引媒體爭相追捧報導。(Source: AP/達志)

概念模型如圖1所示。

雖然「最低限度正義自主使用人工智慧」在技術本質上較為受限，但與「最高限度正義自主使用人工智慧」相比，它在許多特定情況下可能會產生比較符合道德期待的結果。「最低限度正義自主使用人工智慧」不會對受保護的個人或基礎設施採取主動致命或非致命的行動；相較之下，「最高限度正義自主使用人工智慧」需要將規範價值編成碼，並典化成一套套的規則，而且還需要應用複雜且可能不甚完美的機械邏輯，來解釋各式各樣的輸入信息。這種更複雜的演算道德——雖然在某些情況下可能更合乎期待——卻更有可能主動造成致命性的錯誤，特別是在不同利益間操弄衝突的時候。

體認到前述的困境，本文主張開發「最低限度正義自主使用人工智慧」系統以善盡對全人類的基本道德義務乃是基於道德正當性。這種嵌入機器係由設計、工程與作業環境三者間

互相調解的道德主體，甚至要比「最高限度正義自主使用人工智慧」系統更貼近人類。開發「最高限度正義自主使用人工智慧」是超義務行為，也就是它在某些特定情況下能產生道德上的益處，但卻不一定合乎道德所需——甚至可能不合乎道德。

符合最低限度正義人工智慧是「雙邊下注」

有些不喜歡「最低限度正義自主使用人工智慧」者可能主張，當未來支撐「最高限度正義自主使用人工智慧」的人工智慧變更強，大家脫離了以規範為基礎的基本人工神經網路系統，而更趨近於通用人工智慧(Artificial General Intelligence, AGI)時，「最低限度正義自主使用人工智慧」的道德期待值將會減低。他們也主張資源應該都集中於開發符合最高道德的機器人。坦白說，近年來在同性質的不同領域有一些成功演算案例。AlphaGo人機圍棋和Libratus-Much人機撲克牌大戰機器獲勝的立基，就在於演算法持續不斷的發展，而這些成功的案例亦引起許多人注意。[9] 這兩個系統分別與最頂尖圍棋手及撲克牌玩家比賽並獲得優勝，讓這些佼佼者更深切體認到棋、牌藝是他們一輩子都必須努力鑽研的課題。這些初步成果已吸引媒體爭相追捧報導，也讓媒體界對人工智慧發展的潛在機會和相關隱憂突然大感興趣。這些報導並非全然正確，有些甚至造成大眾對人工智慧負面的觀感，引發同前所述「阻止殺手機器人運動」進而提出的反烏托邦觀點。

可想而知的是，不久的將來將出現超級智慧——而實現通用人工智慧的時間預測會在二十到三十年後——報導這類成功故事，卻對近來人工智慧的多次失敗案例避而不談。很多失敗係因商業及機密原因而未予披露，其中一例是微軟的人工智慧泰伊聊天機器人(Tay AI Bot)，它是一款機械學習的聊天機器人，可從與數位使用者的互動中自我學習。在操作一段時間後，泰伊聊天機器人發展出具有強烈性別以及種族歧視的自我人格，迫使微軟最終從服務項目中撤出這款機器人。[10] 臉書也遇過類似問題，它的人工智慧訊息聊天機器人(chatbot)也發展出負面特性。[11] 而且許多具備自駕功能的車輛，也因相關系統無法處置即時情境，以及掛有品質保證的系統無法將此類事件納入因素據以應變，因而引發了多場行車事故。

人工智慧的成功奠基於複雜人工神經網路上，而人工神經網路目前仍存在著許多難以解決的問題。這些由下而上的系統在受控制環境中能夠學得很好，在根據資料結構及其中相關性所建立的腳本中亦能輕易地超越人類，但這些系統在更開放環境裡，像是道路系統和交戰區，都比不上人類由上而下的運思能力。這一類系統在這些環境中存在風險，因為它們需要嚴格恪遵法律和規定。對這些系統提問、解讀、說明、監督及控制存有難度，因為深入學習系統無法輕易自我判斷。[12]

另一個重點是，當更為直覺且也因而更不需解釋的系統開始廣泛運作時，可能無法像以前一樣輕易回復成較早期階段的系統，因為操作者已變得更為依賴系統去做出艱困的決

定。其中風險在於操作者自身的道德決策技能可能已隨著時間而逐漸退化。[13] 倘若此類系統正投入於武裝衝突所需之關鍵任務作戰中，萬一失敗，整個系統可能崩潰，並帶來毀滅性後果。

此外，還有關於功能複雜性，以及得以在無法通訊時還能獨立運作的移動系統，其所會面臨到的實際計算限制問題。通用人工智慧等級系統所需的電腦可能無須微小化，甚或對於所欲目的，也有可能強度不夠或不符成本效益，特別是當自主性武器有時被視為一次性平臺的軍事情境中。[14] 通用人工智慧支持者希望未來電腦處理電源及其他系統元件體積能持續縮小、成本更低，功能也更強大，但卻無法保證未來在沒有量子計算領域優勢下的摩爾定律，能否繼續取得大幅的進展。

此時此刻，不論通用人工智慧最終能否開花結果，「最高限度正義自主使用人工智慧」似乎是產生一個不保證可能後果的遠程目標。相反地，「最低限度正義自主使用人工智慧」系統設法保證人工智慧夠發揮那些明顯且無爭議的好處，相關風險則受一般軍事攻擊定位流程控制。決策者現在就必須採取行動，破除那些可能不會發生的浮誇想像，並且利用現有科技取得正面結果。

執行

《國際人道法》(International Humanitarian Law)第36條：「在研究、開發、取得或採用一項新武器、手段或戰鬥方法時，締約方有義務確定在某些或所有情況下，是否會被此規約或適用於締約方的任何其他國際法規所禁止這些行為。」[15] 1987年，針對此條文的評論更進一步指出各國不但需審查新武器，也需審查任何受改造而改變其功能的現有武器，或進一步審查已通過法律核准但接著又改造過的武器。[16] 因此，嵌入「最低限度正義自主使用人工智慧」的武器將需經過第36條的相關審查。

符合第36條慣例評估作法包括武器的技術描述及性能，並假定由人類評估及決定武器用

(上圖)

對手2D偽裝成磨損的停止路標，將迴旋神經網路(一種深度前饋的人工神經網路，通常大部分被應用在分析視覺影像)應用在智慧安全車輛實驗室(Laboratory for Intelligent and Safe Automobiles)的路標資料庫上，能百分之百正確無誤將每則標示分類成每小時45哩的限制標誌。

(下圖)

使用偵測器後配合分類器——將沒分類的物件圖解成不同分類——百分之百會失敗，每次都把這些標誌辨識成停止標誌。

參戰者可能打算透過類似紅十字或紅新月的標誌，來設法欺騙「最低限度正義自主使用人工智慧」系統以謀求自保，並藉此避開一個合法攻擊。(Source: AP/達志)

法。[17] 在第36條規範下，基於舊有的人類決策功能很明顯被抽離自武器技術機能評估的情況下，人工智慧於是成為評估上的挑戰。評估方法需要延伸到「最低限度正義自主使用人工智慧」的內嵌決策以及行動能力。

雖然第36條刻意規避提及制定之法源，但它可能是為了符合國際紅十字會──以及全人類的利益──於此特定情況下所做出之決議。試想在諸多國際條約中首次提及新武器法律審查的需求。[18] 被視為是第36條前身的《聖彼得堡宣言》(Saint Petersburg Declaration)有著宏觀的視野：「渠等為了維持建立之原則，締約方或同意方保留未來因科學進步，促成軍備改良，並且將戰爭需要和人類法律進行調和。」[19] 武器和自主系統中的「最低限度正義自主使用人工智慧」正適用此項提議。它能夠藉由內嵌式武器性能來辨識受保護物體，然後不對其展開攻擊，因而更符合人道主義的標準。

在第36條條文下，對於共享這套達標的技術資料及演算法則，將可降低執行成本，而且還能為系統尋得反制之法，以提升系統安全。

人道主義反反制作為

批評者可能主張參戰者恐開發出破壞「最低限度正義自主使用人工智慧」武器和自主系統所涉及人道主義成效的各種反制作為。然而，反制「最低限度正義自主使用人工智慧」違反人道主義甚至可能違法。因

此，吾等必須考慮所有會降級、損壞、摧毀或欺騙「最低限度正義自主使用人工智慧」性能的反制手段，以鞏固攻擊目標系統的安全性。

降級、損壞或摧毀：吾等期待那些想要阻止「最低限度正義自主使用人工智慧」達成特定任務，而企圖摧毀或損壞武器性能的敵人，能合法的受到反制。此類反制作為將包括攻擊武器尋標器或其他手段。這類攻擊可能減降、損壞，或摧毀「最低限度正義自主使用人工智慧」的性能。假如該行動是出於自衛，那它就非大家預期之人道主義對象會做出的行為，因此「最低限度正義自主使用人工智慧」的功能無論如何也不需要了。

如果鎖定「最低限度正義自主使用人工智慧」的減降、損壞或摧毀行動是為了造成人道主義的浩劫，那這就是一種犯罪行為了。然而在事情發生時要注意的前提，就是採取行動前必須先忽略目標是否合乎法律規範，因為攻擊行為本身才應該是吾等的主要關注點。

欺騙：參戰者可能只是想要透過類似紅十字(Red Cross)或紅新月(Red Crescent)標誌來設法欺騙「最低限度正義自主使用人工智慧」性能以自保，藉此避開一個合法攻擊。這是《國際人道法》第37條所提及的一種背信行為。然而，如此作為也有反制之法，只要與紅十字會交叉比對這些地點就能找出異常之處。此外，紅十字會標誌是一個鮮明標記，因此透過廣域監視就有助於查探出這類欺騙行為。此外，也因為這個理由，吾等界定出「最低限度正義自主使用人工智慧」道德武器只會對受保護物體或行為「預料外」之存在做出回應。此回應是在攻擊鎖定過程中所下的決定，而且自外於道德武器(如圖1所示)。也會製作究責紀錄，隨後還有行動審查。

人工神經網路是性能最高的物體辨識系統；然而，高性能在程度上就容易出現弱點。研究學者已發現與穩定性有關的一個現象，那就是輸入時若有小擾動，而人類無法感知如此非隨機擾動，但卻可顯示在測試影像上，並造成判斷突然改變。[20] 自此就有許多心力投入這些「對抗樣本」。[21] 在為數眾多且形式不同的攻擊中存著一系列反制。與「最低限度正義自主使用人工智慧」有關的對抗樣本次類別，就是那些可應用在2D或3D實體上，以改變其在機器接收到的外貌對抗樣本。最近對抗演算法已被用來製作「偽裝顏料」，甚至還造成一般深度人工網路分類器判讀錯誤的3D列印產品。[22] 其他擔憂還包括可能會在物體上塗上紅十字會的標誌，此標誌是武器搜尋器可認出，但肉眼卻看不到，還有如圖2所示的兩個圖示，在受保護標誌上繪製看似久經摧殘而磨損的圖型，肉眼不會注意到，但卻會造成演算結果無法辨識此標誌──在這個情況下，一個交通停止路標當然就很像紅十字會標誌了。

相較於這些網路媒體追捧報導的結果，研究學者在相同實驗設定中已展現出「無誤」結果，一如上圖路標和多次現場測試。這些研究學者解釋先前團隊已陷於「偵測器」(一種深度前饋的人工神經網路，最常被應用來分析視覺影像)的迷陣中(如Faster以區域為基礎的迴旋人工神經網路[R-CNN]，搭配將未分類物件圖解成不同類別的分類器)。[23] 這些實驗早期使用的方法因為流水線(pipeline)問題而產出錯誤，包括完美手動截圖(替

代不在現場之偵測器），以及在使用分類器前就更改圖像比例。在實驗室環境外，渠等已構思出一種能勝任真實世界各種角度、範圍以及光線狀況的通用偵測器，但仍需要進一步研究。

開放全球取得「最低限度正義自主使用人工智慧」編碼和資料，如紅十字會圖像和「戶外」影像場景，對確保這些技術在真實狀況與建築中，都將被持續測試與強化有頗大幫助。讓全世界取得「最低限度正義自主使用人工智慧」演算法和資料將加快其執行速度，可為原本負擔不起此項創新費用的國家，提供一種成本更低廉的解決方案，並且對拒絕使用此資源的國防工業公司施加道德壓力。

結論

筆者提出「最低限度正義自主使用人工智慧」

註釋

1. *See* Article 36. International Committee of the Red Cross,"Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977,"accessed 24 October 2018, https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/7c4d08d9b287a42141256739003e636b/f6c8b9fee14a77fdc125641e0052b079.
2. "The Solution," *Campaign to Stop Killer Robots*, 2018, https://www.stopkillerrobots.org/the-solution/.
3. Damminda Alahakoon et al, "RE: An International Ban on the Weaponization of Artificial Intelligence (AI)," website of Prof. Toby Walsh, University of New South Wales, 2 November 2017, https://www.cse.unsw.edu.au/~tw/ letter.pdf.
4. "Keep Killer Robots Science Fiction," *Ban Lethal Autonomous Weapons* (vlog), 10 November 2017, https://au-tonomousweapons.org slaughterbots/.
5. The United States, of course, never ratified the Ottawa treaty but rather chose a technological solution to end the us of persistent land mines—land mines that can be set to self-destruct or deactivate after a predefined time period—making them considerably less problematic when used in clearly demarcated and confined zones such as the Korean Demilitarised Zone. For information *see* Lorraine Boissoneault,"The Historic Innovation of Land Mines—And Why We've Struggled to Get Rid of Them,"*Smithsonian*, accessed October 24, 2018, https://www.smithsonianmag.com/innovation/historic-innovation-land-minesand-why-weve-struggled-get-rid-them-180962276/.
6. Patrick Chisan Hew, "Artificial Moral Agents Are Infeasible with Foreseeable Technologies," *Ethics and Information Technology; Dordrecht* 16, no. 3 (September 2014): 197–206, http://doi.org/10.1007/s10676-014-9345-6.
7. Ronald C. Arkin, Patrick Ulam, and Brittany Duncan, "An Ethical Governor for Constraining Lethal Action in an Autonomous System" (technical report, Fort Belvoir, VA: Defense Technical Information Center, 1 January 2009), https://doi.org/10.21236/ADA493563.
8. Ibid.
9. James O'Malley "The 10 Most Important Breakthroughs in Artificial Intelligence," *TechRadar*, a10 January 2018, https://www.techradar.com/news/the-10-most-important-breakthroughs-in-artificial-intelligence.
10. John West, "Microsoft's Disastrous Tay Experiment Shows the Hidden Dangers of AI," *Quartz*, 2 April 2016, https://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai/.
11. Alex Hern, "Please, Facebook, Don't Make Me Speak to Your Awful Chatbots," *Guardian*, 29 April 2016,

的案例說明可能在今日世界中做出拯救人命的決策。最大的希望在於將原本投注在那些因猜測性恐懼而造成抗爭運動的大批資源，能在未來的某一日移轉至因未配置「最低限度正義自主使用人工智慧」武器系統而造成苦難的人員身上。

作者簡介

Jai Galliott博士目前負責帶領新南威爾斯大學澳大利亞國防研究院的國防安全科技價值小組。他曾任澳大利亞前皇家海軍軍官及軍事研究員，目前亦為西點軍校現代戰爭學院客座研究員，並於牛津大學科技及全球事務中心擔任訪問研究員。

Jason Scholz教授擁有南澳大學電子工程學士學位，並在阿德萊德大學取得電機工程博士學位。他現任「可信賴自主系統國防合作研究中心」首席科學家、澳大利亞國科會技術顧問以及澳大利亞公司董事學會研究員，並於新南威爾斯大學澳大利亞國防研究院擔任教授。

Reprint from *Journal of Indo-Pacific Affairs* with permission.

https://www.theguardian.com/technology/2016/apr/29/please-facebook-dont-make-me-speak-to-your-awful-chatbots.

12. Martin Ciupa, "Is AI in Jeopardy? The Need to Under Promise and Over Deliver—The Case for Really Useful Machine Learning," in *Computer Science & Information Technology (CS & IT)* (Fourth International Conference on Computer Science and Information Technology, Academy & Industry Research Collaboration Center [AIRCC], 2017), 59–70, https://doi.org/10.5121/csit.2017.70407.
13. Jai Galliott, "The Limits of Robotic Solutions to Human Challenges in the Land Domain," *Defence Studies* 17, no. 4 (2 October 2017): 327–45, https://doi.org/10.1080/14702436.2017.1333890.
14. Ibid.
15. Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, https://ihl-databases.icrc.org/ihl/WebART/470-750045?OpenDocument.
16. Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, Commentary of 1987, "New Weapons," 421, https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=F095453E41336B76C12563 CD00432AA1.
17. "A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977," *International Review of the Red Cross* 88, no. 864 (December 2006): 931, https://doi.org/10.1017/S1816383107000938.
18. "Declaration Renouncing the Use, in Time of War, of Certain Explosive Projectiles. Saint Petersburg, 29 November/11 December 1868," accessed 24 October 2018, http://www.gwpda.org/1914m/gene68.html.
19. Ibid.
20. Christian Szegedy et al., "Intriguing Properties of Neural Networks," *Computing Research Repository*, 19 February 2014, https://arxiv.org/pdf/1312.6199.pdf.
21. Naveed Akhtar and Ajmal Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Sur-vey," *IEEE Access* 6 (2 January 2018), 14410–30, https://arxiv.org/pdf/1801.00553.pdf.
22. Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification" (paper, 2018 Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 18–22 June 2018), https://arxiv.org/pdf/1707.08945.pdf.
23. Jiajun Lu et al., "Standard Detectors Aren't (Currently) Fooled by Physical Adversarial Stop Signs," 26 October 2017, https://arxiv.org/pdf/1710.03337.pdf